Procesamiento del Lenguaje Natural, núm. 28 (2002), pp. 89-98 recibido 15-3-2002; aceptado 2-5-2002

Utilización de pasajes de tamaño variable, para mejorar el proceso de recuperación de información

Fernando Llopis
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
llopis@dlsi.ua.es

Antonio Ferrández
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
antonio@dlsi.ua.es

José Luis Vicedo
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
vicedo@dlsi.ua.es

Resumen: Trabajos previos demuestran que la utilización de fragmentos de documentos como unidad básica de información, para calcular la relevancia de un documento con respecto a una pregunta, mejora sensiblemente los resultados de los sistemas de recuperación de información. Sin embargo, no se ha llegado a un consenso acerca de cómo definir esos fragmentos de texto (o párrafos) de forma que el sistema alcance un comportamiento óptimo. El presente artículo presenta un sistema de recuperación de información, basado en la definición de pasajes de tamaño variable. Cada pasaje está formado por un número determinado de las frases que forman el documento. El número de frases seleccionadas para cada pasaje dependerá de la localización de las palabras de la pregunta en cada documento. La evaluación realizada permite comparar el rendimiento de este modelo con un sistema estándar de recuperación de documentos, así como con otras propuestas que utilizan diferentes métodos de definición de pasajes.
Palabras clave: Recuperación de Información. Recuperación por Pasajes. Búsqueda de Respuestas

Abstract: Previous works show that the use of fragments of documents as the basic unit of information to calculate the relevance of a document with regard to a query, improves the results of Information Retrieval systems. However, it has not been agreed how these fragments of texts should be obtained in order to obtain optimum results. This paper presents an Information Retrieval system that is based on the definition of passages of variable size. Each passage is formed by a number of sentences that formed the document. The number of sentences by each passage will depend on the position of the query words in each document. The accomplished evaluation allows comparing the performance of this model with a standard Information Retrieval system, as well as with other Passage Retrieval systems.
Keywords: Information Retrieval. Passage Retrieval. Question Answering.

1 Introducción

Dada una colección de documentos y una determinada pregunta, los sistemas de Recuperación de Información (Information Retrieval, IR) tienen como principal objetivo el de ordenar los documentos de dicha colección en función de su relevancia a la pregunta. Por otro lado los sistemas de Búsqueda de Respuestas (Question Answering, QA) intentan mejorar la salida de los sistemas de IR, devolviendo pequeños fragmentos de texto de la colección que contengan la respuesta a la pregunta realizada (ver Figura 1).

Los sistemas de QA, utilizan tanto técnicas de IR como de procesamiento de lenguaje natural para esa localización de la respuesta. Dado que estas técnicas suelen tener un coste temporal de ejecución elevado, es difícil aplicarlas de forma directa a colecciones de gran tamaño.

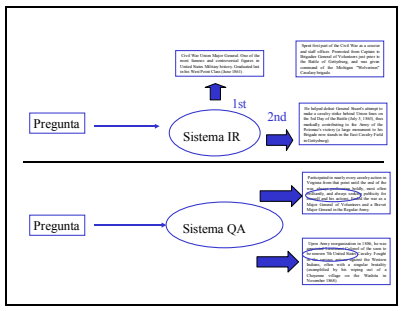

Figura 1: Recuperación de Información y Búsqueda de Respuestas

Debido a esto, muchos sistemas de QA aplican en primer lugar un sistema de IR de forma previa (ver Figura 2) que permita determinar cuáles son los documentos más relevantes, para bien tratar los primeros (Litkowski, 2000) o seleccionar los párrafos más relevantes (Harabagiu et al., 2000).

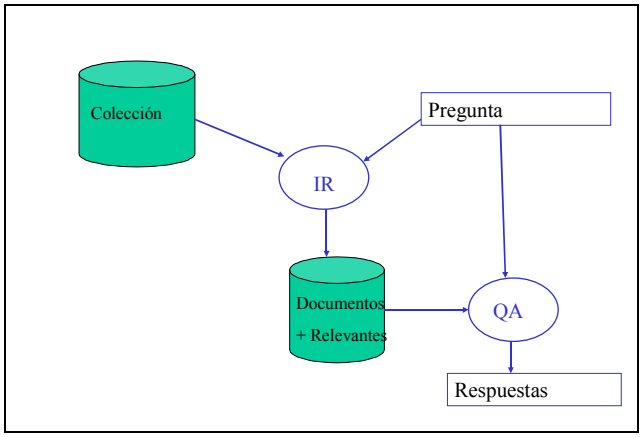

Figura 2. Utilización de la salida de un sistema IR en un sistema QA

Existen diferentes técnicas para determinar la relevancia o similitud de un documento con respecto a una pregunta, pero la mayoría de ellas se basan fundamentalmente en la aparición de los términos que forman la pregunta en el documento completo.

Una alternativa a este método, es realizar esta búsqueda en fragmentos contiguos de texto, a los que denominaremos pasajes. Esto permite no solo valorar el hecho de que los términos de la pregunta aparezcan en el documento, sino que, además, se valora la proximidad en la que aparezcan. Adicionalmente ofrece la ventaja de indicar no solo qué documento es relevante, sino que permite localizar con mayor precisión, qué parte del documento (o pasaje) es realmente relevante. Este aspecto es de gran utilidad por ejemplo, cuando la salida del sistema de IR es utilizada por un sistema de QA, ya que limita considerablemente la cantidad de texto que este sistema debe procesar.

La propuesta de sistema de IR que se presenta en este artículo, denominado sistema IR-n, se engloba dentro de los sistemas de recuperación que se basan en pasajes (Passage Retrieval, PR). El sistema IR-n define los pasajes seleccionando un número fijo de frases del documento. Este sistema fue presentado en la última edición del Cross-Languaje Evaluation Forum, CLEF-2001 (Llopis y Vicedo, 2001), como sistema de IR. También fue utilizado en la TExt Retrieval Conference, TREC-10, (Vicedo, Ferrández y Llopis, 2001), dentro de la tarea de QA, como sistema que seleccionara los pasajes más relevantes, para posteriormente aplicar un sistema de QA integrado.

Nuestras últimas investigaciones se han dirigido a dotar al sistema de la posibilidad de definir los pasajes en base a un número variable de frases, y a estudiar el rendimiento de esta propuesta. Esta, es presentada en este artículo y se evalúan y comparan los resultados respecto a la propuesta anterior basada en pasajes de tamaño fijo, así como con otras propuestas de sistemas de PR o IR basadas en el documento completo.

En la siguiente sección se presentarán los antecedentes de los sistemas de IR y PR. A continuación, en la sección 3 se detallará la propuesta de este artículo, para luego en la sección 4 realizar la comparativa de los resultados del sistema con otros sistemas. Se finalizará en el apartado 5 con las conclusiones y el trabajo en curso que se está realizando.

2 Del estudio del documento completo a los sistemas de recuperación por pasajes

En la presente sección se comentará en primer lugar los problemas que tienen los sistemas de IR que estudian el documento completo, para luego definir las características principales y clasificación de los sistemas de PR.

2.1 Problemáticas en la recuperación por el documento completo

Para determinar el grado de relevancia de un documento con respecto a una pregunta, se suelen buscar en el primero si aparecen los términos que forman la pregunta.

Adicionalmente se suele valorar también:

1. El número de veces que aparecen los términos de la pregunta en el documento. A mayor número de apariciones mayor valoración para el documento.
2. El peso de cada término. A mayor número de documentos donde aparezca el término menor valoración del término.
3. El tamaño del documento (medido en bytes o palabras). Aplicándose unas

medidas de normalización, que permitan valorar documentos de diferente tamaño.

Las medidas más conocidas para el cálculo de similitud entre documentos y preguntas son tres, la medida del *coseno* (Salton, 1989) el *coseno pivotado* (Singhal, Buckley y Mitra, 1996). y el sistema *okapi* (Roberston y Walker, 1994).

Los mayores problemas de los sistemas de RI que se basan en el estudio del documento completo son:

1. Determinan qué documento es relevante o no, pero no especifican la parte del documento que realmente es relevante a la pregunta. Un documento titulado "Biografía de Felipe II", será relevante con respecto a una pregunta del tipo "El nacimiento de Felipe II", pero solo una parte de dicho documento será el que hace referencia concretamente a dicha pregunta. Por ejemplo, este aspecto es de fundamental importancia cuando la salida del sistema de IR es la entrada de un sistema de QA.
2. No tienen en cuenta el concepto de proximidad en la aparición de los términos de la pregunta en el documento. En la Figura 3 se puede ver cómo 2 documentos que serían valorados de la misma forma pueden no ser realmente igual de relevantes.
3. No está suficientemente claro la forma de valorar documentos de diferentes tamaños.

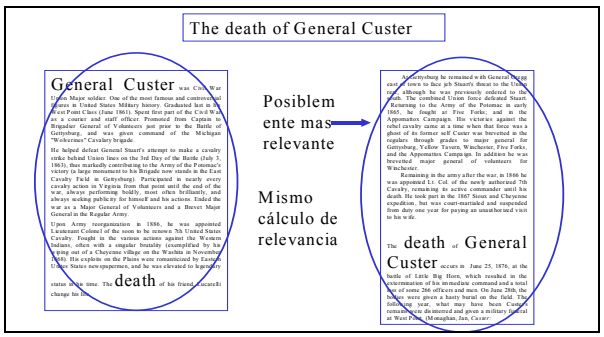

Figura 3. Problemática de IR con documento completo

2.2 Los sistemas de recuperación por pasajes

Para paliar estos problemas han aparecido otras propuestas que se basan en la aplicación de medidas de relevancia a pequeños fragmentos (pasajes) de los documentos. Estos sistemas se denominan Recuperación por Pasajes o Passage Retrieval (PR). Estos sistemas se basan en la división previa de cada documento en fragmentos, para luego aplicar las medidas de relevancia a cada uno de estos fragmentos (ver Figura 4).

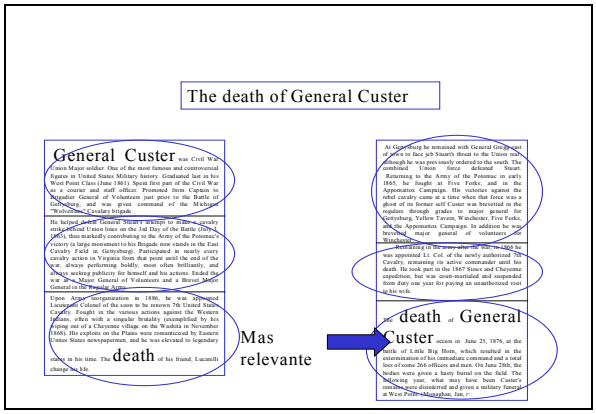

Figura 4. Cálculo de relevancia por pasajes

Trabajos previos (Callan, 1994) (KaszKiel y Zobel, 2001) indican que los sistemas de PR mejoran a los sistemas de IR que se basan en el documento completo.

Fundamentalmente los modelos propuestos de PR se diferencian entre ellos en la forma en la que se definen los pasajes en los que se divide el documento. Una clasificación generalmente aceptada es la definida en (Callan, 1994), donde se diferencian los sistemas de PR en modelos basados en el discurso, modelos semánticos y modelos de ventana.

Los modelos del discurso (Namba, 2000)(Salton, Allan y Buckley, 1993) utilizan las propiedades de estructura del documento, tales como frases, marcas de párrafo, marcas HTML, para definir los párrafos.

Los modelos semánticos se basan en la aparición de tópicos en el documento para definir los pasajes (Hearst y Plaunt, 1993).

Los modelos de ventana dividen los documentos en pasajes de tamaño fijo. Dentro de los modelos de ventana se hace una subclasificación adicional en (KaszKiel y Zobel, 2001), donde se diferencian aquellos que utilizan la estructura del documento en el momento de definir los pasajes o aquellos que no la utilizan. Los primeros se basan en la

unión de párrafos de tamaño pequeño o manteniendo los párrafos de cierto tamaño. En (Zobel et al. 1995) se utilizan medidas basadas en bytes (definiendo los pasajes alrededor de los 2 kBytes, denominando al sistema *PAGES*) y en el modelo *Bounded Paragraphs* (Callan 1994) se definen los pasajes en función del número de palabras que contienen los párrafos (definiendo los pasajes entre 50 y 200 palabras). Los modelos de ventana que no se basan en la estructura del documento pueden iniciar los pasajes en cualquier palabra del texto con más o menos limitaciones. Dentro de estas propuestas destacan las de *Sliding Windows* (Callan 94) y *Arbitrary Passages* (Kaszkiel y Zobel 1997).

Parece coherente pensar que los modelos del discurso van a ser más efectivos ya que utilizan la propia estructura del documento definida por el autor del mismo. No obstante los mayores problemas de utilización de estos modelos, es que frecuentemente los párrafos pueden ser utilizados más por motivos visuales que de contenido y que pueden depender excesivamente del estilo utilizado por los autores del documento. Una problemática añadida es que generan una colección muy heterogénea en cuanto a tamaños de pasajes. Los modelos de ventana por otro lado generan un conjunto de pasajes mucho más homogénea con respecto al tamaño y suelen ser mucho más fáciles de construir que con los otros modelos. Dentro de los modelos de ventana, aquellos que pueden empezar en cualquier parte del documento pueden perder cierta estructura y pueden ser no tan recomendables para presentar al usuario el pasaje más relevante o bien cuando son utilizados por otros sistemas que requieren de cierta estructura en el pasaje, como los sistemas de QA, aunque puede ser de gran interés su aplicación a colecciones de texto en las que no sea sencillo determinar su estructura como podría ocurrir en los diálogos.

3 El sistema IR-n

El sistema IR-n (Llopis y Vicedo 2001) se engloba dentro de los modelos de ventana que utilizan la estructura del documento, utilizando para definir los pasajes un número fijo de frases. En esta sección de especificará en primer lugar la arquitectura del sistema IR-n, a continuación se indica el método que aplica para el cálculo de relevancia y finalmente se define la propuesta de añadir al sistema la posibilidad de utilizar pasajes de tamaño variable.

3.1 Arquitectura del sistema IR-n

Los principales módulos que componen el sistema IR-n son 3, Indexador, Preprocesador y Buscador. La arquitectura del sistema puede verse de forma gráfica en la Figura 5.

El módulo indexador es el responsable de organizar la información contenida en la colección de documentos para facilitar su acceso.

El módulo preprocesador es el responsable de preparar los términos de la pregunta para su posterior búsqueda. Este módulo también es el encargado de determinar y realizar la expansión de la pregunta.

El módulo buscador es el responsable de calcular la relevancia de cada documento con respecto a la pregunta. El modelo de relevancia que utiliza se indica en la siguiente subsección.

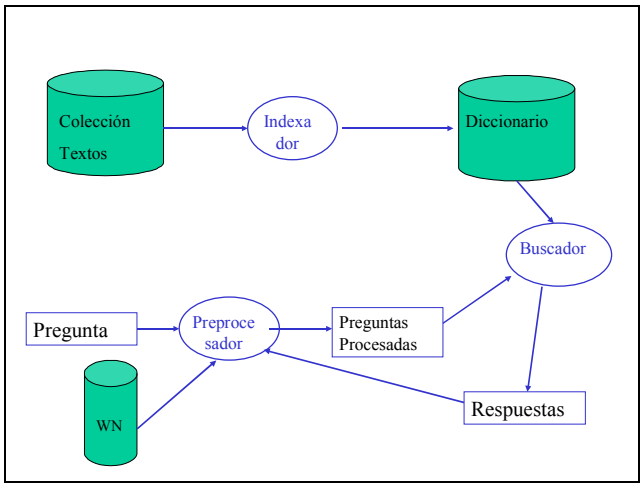

Figura 5: Arquitectura del sistema IR-n

3.2 Modelo de relevancia en el sistema IR-n

Las principales características en las que se basa el sistema IR-n para el cálculo de relevancia entre documentos y preguntas son las siguientes:

1) Un documento se divide en pasajes formados por un número determinado de frases. Esto se debe a que consideramos que una frase suele representar una idea dentro del documento, mientras los párrafos pueden ser utilizados más por motivos visuales. Además,

las frases son unidades completas que permiten un tratamiento posterior por otro sistema, como es el caso de un sistema de QA, mientras que los pasajes que empiezan en cualquier palabra del documento pueden perder parte de su estructura.

2) El número de frases que definen el tamaño del pasaje depende del tipo de la pregunta y de la colección de documentos. Actualmente hemos obtenido los mejores resultados utilizando entre 15 y 20 frases.

3) El sistema utiliza ventanas que se solapan unas sobre otras. El grado de solapamiento es una frase. Es decir, si el número de frases que forman el pasaje es 15, entonces el primer pasaje estaría formado por las frases de la 1 a la 15, el segundo por las frases de la 2ª a la 16º y así sucesivamente (ver Figura 6). Experimentalmente (Llopis, Ferrández y Vicedo, 2002) se demostró que el definir este tipo de solapamiento de los pasajes obtenía mejores resultados que cuando no se utilizaban pasajes solapados o se utilizaba un grado de solapamiento mayor. Utilizamos ventanas definidas de este tipo ya que experimentalmente parece evidente que esta definición de los pasajes incrementa el tiempo de respuesta del sistema. No obstante este incremento no es excesivo ya que realmente el sistema IR-n considera que el primer pasaje empieza en la primera frase donde aparece uno de los términos de la pregunta, y el último pasaje finaliza en la última frase del documento donde aparece un término de la pregunta.

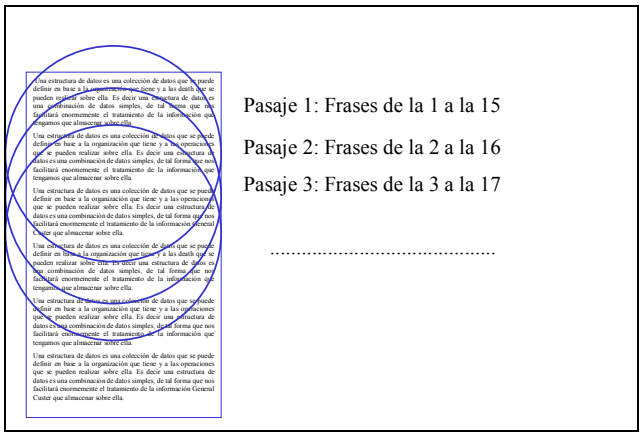

Figura 6: Definición de pasajes en el sistema IR-n

4) La medida de similitud que se aplica a cada pasaje es la mostrada en (1). Dado una pregunta *q* y un pasaje *p*.

$$\text{Similitud del párrafo} = \sum_{t \in p \wedge d} W_{p,t} * W_{q,t} \quad (1)$$

donde :

$W_{p,t} = log_e(f_{p,t} + 1)$. Siendo $f_{p,t}$ el número de apariciones del término *t* en el pasaje *p*.

$W_{q,t} = log_e(f_{q,t} + 1) * idf$. Siendo $f_{q,t}$ el número de apariciones del termino *t* en la pregunta *q*.

$idf = log_e(N/f_t + 1)$. Siendo *N* el número de documentos de la colección, y f_t es el número de documentos diferentes en los que aparece el término *t*.

Cabe indicar que esta medida es similar a la del coseno extrapolándola a pasajes, pero no utiliza normalización alguna respecto al tamaño de los pasajes, ya que el sistema considera la frase como una unidad y todos los pasajes están formados por el mismo número de frases.

5) El sistema ha sido desarrollado en C++ sobre una plataforma Linux. El sistema no requiere para funcionar software adicional ni elevados requerimientos hardware.

3.3 Características del sistema IR-n con pasajes variables

Con el objeto de intentar mejorar la aproximación comentada en el apartado previo, se propone en este artículo una definición de los pasajes más flexible. Parece evidente que un pasaje será más relevante que otro, si a pesar de tener la misma medida de similitud, las apariciones de los términos de la pregunta se agrupan en una serie determinada de frases consecutivas. Para ello se propone la posibilidad de valorar esa agrupación. La propuesta IR-n con pasajes de tamaño variable se basa en recortar en cada pasaje las primeras y últimas frases que no contengan ningún término de la pregunta. Así un pasaje formado por 15 frases, de las cuales la primera y última que contienen algún término de la pregunta son la 4ª y la 10ª estaría formada por las 7 frases que se hallan entre dichas dos.

Esta definición de los pasajes hace aconsejable que se añada un componente de normalización que incremente la puntuación a aquellos pasajes que engloban los términos de la pregunta en un menor número de frases consecutivas. Esta normalización debe ser lo suficientemente suavizada, para que no obtengan mejor medida de relevancia pasajes

de pocas frases que contengan pocos términos, sobre pasajes de mayor tamaño en los que aparecen gran cantidad de términos de la pregunta. Así al cálculo de relevancia definido en el sistema IR-n se le añade un factor de normalización de forma logarítmica que incrementa la medida de relevancia de un pasaje cuanto menor es el número de frases que lo forman el pasaje. La medida propuesta se detalla en (2).

$$S_v = S_f * (1-x) + (S_f * x / \sqrt{\log_e(n)}) \quad (2)$$

Siendo:

S_f la medida de similitud que se aplica en el sistema IR-n con pasajes de tamaño fijo.

n el número de frases consecutivas de un pasaje que contienen los términos de la pregunta que aparecen en dicho pasaje.

x Un parámetro que determina la importancia a aplicar del tamaño seleccionado. Cuando mayor es, más valor se le da a los pasajes de menor tamaño. Se le ha aplicado una medida de suavizado de forma logarítmica, Actualmente en las pruebas realizadas se le ha fijado un valor de 0,5.

4 Evaluación

En este apartado se presentan los resultados obtenidos por la propuesta realizada del sistema IR-n con pasajes variables, realizando una comparación con otros sistemas de IR y PR, así como con la versión del sistema IR-n basada en pasajes de tamaño fijo.

4.1 Colección de documentos y preguntas

La colección de preguntas y documentos son las utilizadas para la lengua inglesa en la edición de 2001 del CLEF. La colección de documentos es la de *Los Angeles Times 94*. Esta colección está formada por 113.005 documentos. Esta colección es muy heterogénea con respecto al tamaño de los documentos, el mayor está formado por 807 frases y el más pequeño sólo por 1.

Cada pregunta tiene 3 versiones, título, descripción y narrativa. El título suele estar formado por entre 2 y 4 palabras, la descripción es similar al título y la narrativa detalla en mayor medida la pregunta, estando formado por varias frases. A continuación se puede ver una de las preguntas de la colección:

Titulo: *"Pesticides in Baby Food"*.

Descripción: *"Find reports on pesticides in baby food"*.

Narrativa*: "Relevant documents give infomation on the discovery of pesticides in baby food. They report on different brands, supermarkets, and companies selling baby food which contains pesticides. They also discuss measures against the contamination of baby food by pesticides"*.

Hay que indicar que en todas las preguntas se filtran las palabras irrelevantes, ya sean las stop-words (*of, which, ...*) o las que hacen referencia a la tarea a realizar (relevant documents o find reports).

Se han llevado a cabo 2 experimentos, el primero que utiliza solo el título de la pregunta (al que denominaremos pregunta corta) y el segundo que utiliza título, narrativa y descripción (al que denominaremos pregunta larga).

Las medidas de evaluación utilizadas, son las de precisión y cobertura, que se definen en las fórmulas (3) y (4).

$$\text{Cobertura} = \frac{REL\_EXT}{REL\_COL} \quad (3)$$

$$\text{Precisión} = \frac{REL\_EXT}{DOC\_REC} \quad (4)$$

Siendo

REL_EXT. El número de documentos relevantes recuperados

REL_COL El número de documentos relevantes en la colección

DOC_REC El número de documentos recuperados

4.2 Resultados obtenidos

Para comparar nuestras propuestas se han realizado pruebas sobre la misma colección de documentos y preguntas con diferentes modelos de IR. Además de las propuestas del sistema IR-n con pasajes de tamaño fijo o variable (utilizando en ambos casos pasajes de 15 y 20 frases), se han evaluado un sistema de IR basado en el documento completo, el sistema SMART (Buckley, Allan y Salton, 1994), un conocido sistema de IR basado en el modelo del coseno y 4 modelos de PR, 1 modelo basado en el discurso utilizando los párrafos de los documentos para definir los

pasajes (Salton, Allan y Buckley,1993), un modelo semántico el TextTiling (Hearst y Plaunt, 1993) y dos modelos de ventana Bounded Paragraphs, en adelante BP, (Callan, 1994) y PAGES (Zobel et al.1995), todos ellos comentados en la sección 2.2 del artículo.

En la Tabla 1 se pueden ver los resultados referentes a la precisión que obtienen los diferentes sistemas evaluados a un número de documentos recuperados determinado (5,10,15, 20 30 y 100). Los modeles de ventana obtienen mejoras sensibles sobre el modelo basado en documento completo (SMART) y los modelos de PR basados en el discurso o semánticos. Cabe destacar los resultados del sistema IR-n en esta tabla, entre un 95%, a los 5 documentos recuperados, y un 20& a los 100 documentos recuperados.

En la Tabla 2 se pueden ver los resultados de los diferentes sistemas a nivel de medias de Cobertura y Precisión interpoladas para las preguntas cortas. Aquí se observa que los sistemas de PR, excepto el sistema IR-n ofrecen resultados similares o incluso sensiblemente peores que el sistema SMART. No obstante la nueva propuesta del sistema IR-n obtiene una mejora del 83% tanto sobre el SMART como sobre el mejor de los sistemas de PR evaluados (BP). Además, se obtiene una mejora del 5% sobre el sistema IR-n basado en pasajes de tamaño fijo. En la Figura 7 se puede ver una comparativa de forma gráfica de las dos versiones del sistema IR-n evaluadas, así como del sistema SMART y el mejor de los sistemas de PR evaluados (BP). Del estudio de estas dos tablas, se puede extraer la conclusión de que los sistemas de PR, recuperan más velozmente los documentos relevantes (mejores resultados en Tabla 1), pero no obtienen tan buenos resultados a nivel de medias de cobertura y precisión, debido a que algunas soluciones a las preguntas pueden hallarse en dos de los pasajes definidos en las diferentes propuestas. Esto no ocurre así en el sistema IR-n, ya que este utiliza pasajes construidos de forma solapada, con lo cual se obtienen mejores resultados a costa de un mayor coste de tiempo de ejecución.

En la Tabla 3 se puede ver la misma comparativa, pero en este caso utilizando las preguntas largas. En este caso los resultados de las 2 propuestas del sistema IR-n se igualan, obteniéndose una mejora del 48% sobre el sistema SMART y de un 21% sobre el sistema BP. En la Figura 8 se puede visualizar de forma gráfica estos resultados, comparándose los dos modelos del sistema IR-n evaluado, así como un sistema basado en documentos completo (SMART) y el que mejor resultado ha obtenido de los sistemas de PR evaluados.

	5	**10**	**15**	**20**	**30**	**100**
Smart	0,2298	0,2021	0,1830	0,1649	0,1447	0,0826
Paragraphs	0,1872	0,1660	0,1558	0,1383	0,1248	0,0787
Tiling	0,2298	0,1872	0,1674	0,1543	0,1376	0,0849
Pages	0,2553	0,2532	0,2426	0,2234	0,1943	0,1055
BP	0,2596	0,2404	0,2270	0,2160	0,1837	0,1019
IR-n 15frases	0,5064	0,4043	0,3348	0,2926	0,2397	0,1211
IR-n 20 frases	0,4851	0,4000	0,3376	0,2904	0,2468	0,1189
IR-n 15 frases var.	0,5021	0,4085	0,3376	0,2915	0,2404	0,1245
IR-n 20 Frases var.	0,4936	0,4043	0,3376	0,2947	0,2411	0,1253

Tabla 1: Precisión a los n documentos recuperados, en preguntas cortas

	Smart	**Párrafos**	**Tiling**	**Pages**	**BP**	**IR-n 15 frases**	**IR-n 20 frases**	**IR-n 15 frases var**	**IR-n 20 frases var**
a 0.00	0.5264	0.3907	0.3798	0.4574	0.4697	0.7601	0.7816	0.7699	0.7984
a 0,10	0,4003	0,3299	0,3079	0,3430	0,3977	0,7338	0,7354	0,7440	0,7535
a 0,20	0,3389	0,2785	0,2494	0,3165	0,3556	0,6948	0,6927	0,6993	0,7112
a 0,30	0,3098	0,2283	0,2058	0,2776	0,3209	0,5790	0,5898	0,5821	0,6010
a 0,40	0,2693	0,2008	0,1704	0,2506	0,2823	0,4999	0,4864	0,5067	0,5049
a 0,50	0,2549	0,1896	0,1520	0,2300	0,2561	0,4436	0,4376	0,4667	0,4600
a 0,60	0,1998	0,1211	0,1095	0,1972	0,2079	0,3344	0,3290	0,3540	0,3503
a 0,70	0,1746	0,1044	0,0879	0,1742	0,1766	0,2723	0,2724	0,2877	0,3000
a 0,80	0,1515	0,0867	0,0713	0,1539	0,1577	0,2389	0,2385	0,2582	0,2598
a 0,90	0,1279	0,0686	0,0545	0,1238	0,1290	0,1942	0,2002	0,2067	0,2177
a 1,00	0,1035	0,0386	0,0320	0,0957	0,0817	0,1317	0,1339	0,1421	0,1493
Precisión media	0,2426	0,1720	0,1526	0,2202	0,2437	0,4255	0,4270	0,4385	0,4463

Tabla 2: Medias de Precisión y Cobertura con preguntas cortas

	Smart	Párrafos	Tiling	Pages	BP	IR-n 15 frases	IR-n 20 frases	IR-n 15 frases var	IR-n 20 frases var
a 0,00	0,6646	0,5694	0,4265	0,6544	0,6972	0,8002	0,8173	0,8033	0,8173
a 0,10	0,5663	0,4975	0,3460	0,5988	0,6234	0,7612	0,7853	0,7635	0,7847
a 0,20	0,4603	0,4288	0,2912	0,5475	0,5765	0,7137	0,7229	0,7167	0,7230
a 0,30	0,4125	0,3559	0,2332	0,4733	0,5240	0,6375	0,6343	0,6371	0,6320
a 0,40	0,3890	0,3388	0,2030	0,4339	0,4785	0,5719	0,5648	0,5655	0,5742
a 0,50	0,3402	0,3181	0,1777	0,4017	0,4330	0,5461	0,5325	0,5392	0,5414
a 0,60	0,2819	0,2476	0,1317	0,3635	0,3667	0,4462	0,4381	0,4452	0,4442
a 0,70	0,2545	0,2036	0,1126	0,3183	0,3271	0,3834	0,3910	0,3837	0,3913
a 0,80	0,2282	0,1853	0,0937	0,2933	0,3036	0,3140	0,3174	0,3273	0,3219
a 0,90	0,1982	0,1552	0,0700	0,2713	0,2711	0,2563	0,2465	0,2714	0,2616
a 1,00	0,1614	0,1303	0,0511	0,2086	0,2090	0,1883	0,1844	0,2002	0,1970
Precisión media	0,3416	0,2944	0,1811	0,3960	0,4189	0,4999	0,5014	0,5033	0,5069

Tabla 3: Medias de precisión y cobertura con preguntas largas

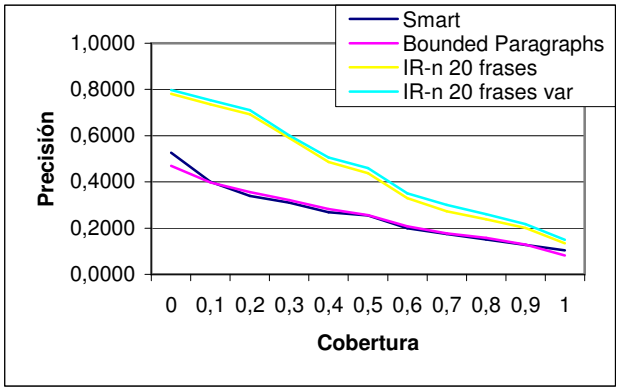

Figura 7: Medias de Precisión y Cobertura en preguntas cortas

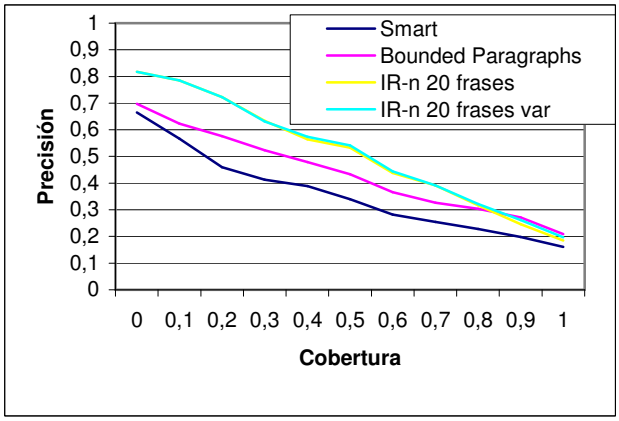

Figura 8: Medias de Precisión y Cobertura en preguntas largas

5 Conclusiones y trabajos futuros

En el presente artículo se ha propuesto un sistema de recuperación de información basado en pasajes, que utiliza las frases como unidad para definir los pasajes. Esta propuesta se basa en una anteriormente realizada (Llopis y Vicedo 2001), con la diferencia que la propuesta actual permite definir los pasajes en base a un número variable de frases, mientras que en la anterior todos los pasajes tenían el mismo número de frases. Al tener que comparar pasajes de diferentes tamaños, se ha incorporado al método de cálculo de las medidas de relevancia, un factor de normalización basado en el número de frases que forma cada pasaje.

Se ha comparado esta nueva propuesta, junto con la original, con diversos sistemas de recuperación por pasajes y uno basado en el estudio del documento completo. Las comparativas realizadas son en primer lugar la precisión a los n documentos, y las medias de precisión y cobertura. Por los resultados obtenidos, se puede indicar que la mayoría de los sistemas basados en pasajes, obtienen mejores resultados de precisión, calculados en base a los primeros documentos recuperados, que los basados en el documento completo. No obstante, a niveles de media de precisión y cobertura, cuando se recuperan gran cantidad de documentos, los resultados son similares.

No obstante, nuestra propuesta el sistema IR-n, obtiene mejores resultados en ambas comparativas. Obteniéndose mejoras, sobre el mejor del resto de sistemas evaluados, de un 95% en cuanto a precisión a los 5 documentos recuperados y de un 20% a los 100 documentos recuperados. A nivel de medias de precisión y cobertura, las mejoras son del 83%

en el caso de utilizar preguntas cortas y entre un 21% y 48 en el caso de las preguntas largas. Además, la nueva propuesta basada en ventanas de tamaño variable, mejora a la anterior en un 5% en las medias de precisión y cobertura, cuando se utilizan preguntas cortas, y resultados similares cuando las preguntas son largas.

Se plantea una serie de trabajos a realizar. El primero de ellos consiste en el estudio de las mejoras que el sistema propuesto, basado en pasajes de tamaño variable, puede suponer cuando se aplica en tareas de QA. Cabe destacar que en la edición del TREC-10 (Vicedo, Ferrández y Llopis 2001) se utilizaron los 200 mejores pasajes, (de 15 frases cada uno), suministrados por el sistema IR-n que utilizaba pasajes de tamaño fijo. Estos pasajes contenían el 95.15% de las respuestas a las preguntas realizadas. Nuestro objetivo es verificar dos cosas fundamentalmente. La primera es comprobar si con esta nueva propuesta se incrementa el número de respuestas incluidas en los pasajes que devuelve el sistema. La segunda es verificar si se facilita la tarea del sistema de QA, disminuyendo el tamaño de los pasajes que éste debe evaluar.

Para finalizar indicar que se pretende dejar disponible en la web del grupo de investigación una versión del sistema IR-n, que se pueda descargar, para que otros grupos de investigación puedan experimentar con el mismo.

Agradecimientos

Este artículo ha sido financiado parcialmente por el Gobierno Español (CICYT) dentro del proyecto número TIC2000-0664-C02-02 y TIC2001-3530-C02-02

Bibliografía

Callan, J. 1994. Passage-Level Evidence in Document Retrieval. *In Proceedings of the 17 th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 302-310, *Dublin,* Ireland.

Harabagiu, S., D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Gîrju, V. Rus y P. Morarescu, 2000. FALCON: Boosting Knowledge for Answer Engines. *In Nineth Text REtrieval Conference,* páginas 479-487, Gaithersburg USA

Hearst, M. y C. Plaunt, 1993. Subtopic structuring for full-length document access. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 59-68, Pittsburgh, PA USA

KaszKiel, M. y J. Zobel, 2001. Effective Ranking with Arbitrary Passages. *Journal of the American Society for Information Science*, Vol 52, No. 4, February, 344-364.

Litkowski, K. 2000. Syntactic Clues and Lexical Resources in Question-Answering *In Nineth Text REtrieval Conference,* páginas 177-188, Gaithersburg USA

Llopis, F. y J. Vicedo, 2001. IR-n system, a passage retrieval system at CLEF 2001 *Working Notes for the Clef 2001*, páginas 115-120. Darmstdt, Germany. Se publicará proximamente en Lecture Notes Computer Science.

Llopis, F., A. Ferrández, y J. Vicedo 2002. Text Segmentation for efficient Information Retrieval. *Third International Conference on Intelligent Text Processing and Computational Linguistics.* Lecture Notes in Computer Science, páginas 373-380, Mexico.

Namba, I . 2000. Fujitsu Laboratories TREC9 Report. *Proceedings of the Nineth Text REtrieval Conference,TREC-9*, pp. 203-208, Gaithersburg,USA.

Roberston, S.E. y S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weight retrieval. *In B.W. Croft & C.J. van Rijsbergen (Eds) Proceedings of the 17th annual ACM-SIGIR conference on research and development in information retrieval,* páginas 232-241*,* Dublin, Ireland

Salton G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, *Addison Wesley Publishing,* New York.

Salton, G., J. Allan, y C. Buckley. 1993. Approaches to passage retrieval in full text information systems. In R Korfhage, *E Rasmussen & P Willet (Eds.) Prodeedings of the 16 th annual international ACM-*

SIGIR conference on research and development in information retrieval, 49-58. Pittsburgh PA USA .

Buckley, C., J. Allan, y G. Salton. 1994. Automatic routing and ad-hoc retrieval using SMART: TREC *En Donna Harman, editor, Proceedings of the Second Text Retrieval Conference TREC-NIST Special Publication , 500-215*.

Singhal, A., C. Buckley, y M. Mitra. 1996. Pivoted document length normalization. *Proceedings of the 19th annual international ACM-SIGIR conference on research and development in information retrieval*..

Vicedo, J. y A. Ferrández, 2000. A semantic approach to Question Answering systems. *In Nineth Text REtrieval Conference, páginas* 440-444. Gaithersburg, USA

Vicedo, J., A. Ferrández y F. Llopis. 2001. University of Alicante al TREC-10. *In Tenth Text REtrieval Conference,* Gaithersburg,USA.

Zobel, J, A. Moffat, R. Wilkinson, y R. Sacks-Davis 1995. Efficient retrieval of partial documents. *Information Processing and Management*, 31(3) páginas 361-377.